An Efficient Web Information Retrieval Model using Machine Learning

SHORT SYNOPSIS

Submitted in partial fulfilment of the requirement of the degree of

DOCTOR OF PHILOSOPHY

To

Manav Rachna International Institute of Research and Studies

(Deemed to be University)

By

Ms. Simple Sharma

Registration number: 18/ Ph.D./106

Under the supervision of

Dr. Seema Verma

Associate Professor, Faculty of Engineering and Technology

FACULTY OF ENGINEERING AND TECHNOLOGY

MANAV RACHNA INTERNATIONAL INSTITUTE OF RESEARCH & STUDIES

SECTOR-43, SURAJKUND –DELHI ROAD, FARIDABAD – 121001

FACULTY OF ENGINEERING AND TECHNOLOGY

Certificate

I hereby declare that the short synopsis "**An Efficient Web Information Retrieval Model using Machine Learning**" being submitted in partial fulfilment of the requirement for the degree of Doctor of Philosophy in Computer Science and Engineering under Faculty of Engineering and Technology of Manav Rachna International Institute of Research & Studies, Faridabad during the academic year 2020-21.

It is a bonafide record of original work carried out under the supervision of Dr. Seema Verma, Associate Professor, CSE, FET, MRIIRS and has not been submitted in full or part to any other university or institute for degree or diploma. Further, it is certified that short synopsis is plagiarism free and proof has been submitted.

Simple Sharma
(18/ Ph.D/106)

Dr Seema Verma
Associate Professor
CSE, FET
MRIIRS, Faridabad

(Chairperson, DRC)

Chairperson, CDP

ABSTRACT

Search Engine is a software program that looks for items and recognizes those items in a database that makes a perfect match to keywords or characters identified by the users, used for finding a particular website on the World Wide Web uniquely. Searching is done with given keywords and a list of webpage documents are reflected based on the matched keywords. Sorting and selection through various webpages is a tedious task in search engine. Search Engine Optimization (SEO) is a method that provides useful information out of billions of webpages with the use of various techniques. Search engines are broadly divided into three categories namely; Crawler-Based Search Engines, Human-Powered Directories and Hybrid Search Engine. Information retrieval in Search Engine is a big challenge. Web mining helps to boost the power of search engine by searching and analysing patterns in web data. Web mining helps SEO professional to analyse information and trends. The mined data can be the text, structured form of data such as lists and tables, images, video and audio. Web mining results in various complicacies while analysing the information pattern. The foremost challenge is the generation of irrelevant data in weblogs due to which complex pre-processing techniques are in use. The complex techniques used in pre-processing consume more time to get the formatted output; hence time delay is the other challenge. Another big challenge is inaccuracy due to previous information stored in the weblogs. The weblogs contain old information based on the user's previous visit to the site, which affects the current requirements. Hence, there is the need to analyse the weblogs for various challenges to optimize the information retrieval from search engines. This research aims to suggest an efficient information retrieval model to improve users' experience after analysing various complexities.

Key words: Search Engine, Crawler, Page-rank, Search Engine Optimization, Relevancy, Web Mining, Web-Usage Mining, Ranking.

CONTENTS

S. No.	Description	Page No.
1	**Introduction……**	**5**
	1.1 Types of Search Engines……...	**5**
	1.1.1 Components of Crawler-Based Engines …….	**6**
	1.1.2. Search Engine Optimization……...	**8**
	1.2 Web mining ……	**11**
	1.2.1 Types of Web Mining …….	**11**
	1.2.2 Web-Usage Mining ……...	**12**
2	**Literature Review…….**	**14**
3	**Description of Broad Area……**	**17**
	3.1 Research Gap ……	**17**
	3.2 Problem Formulation ……	**18**
4	**Objectives of the study…….**	**18**
5	**Methodology……**	**18**
6	**Expected Outcome of the Research…….**	**19**
	References (Harvard Style/APA/MLA)……	**20**
	Plagiarism Report……...	**23**

1. Introduction

Search engine is a software program that helps in searching information on the World Wide Web. Searching through billions of webpages to get the required page is a tedious task. Some of the popular search engines are Google, Yahoo, Bing, Ask, etc. Search engines have to call through all the websites to carry out indexing. They may be slow to index dynamically because the contents are not static. The speed at which the search engine is indexing the pages is increasing day by day but the documents over the Web are increasing at a much faster rate. There are billions of pages that exist in the World Wide Web (WWW). Search engines create a dictionary where size of the dictionary needs to be very large. Today search engines are becoming more efficient and powerful by using various algorithms and techniques to provide the best results demanded by user. As an example, Google has introduced some simple concepts which allows a user to get more relevant search results as compared to other competing search engines like, link popularity, the number of pages on a website, recording pages that point to another webpage, etc. All these concepts increased the visibility of the page and the website effectively communicates with major search engines. Optimizing a page involves editing the contents of the webpage so as to extend its relevancy to identified words. Many researchers have developed numerous search engine optimization strategies, including (Gupta and Anand, 2015), (Liu, Xu and Long, 2017), (Omprakash, 2011), (Chen *et al.,* 2017), (Kumar and Kumar, 2014), (Singh, 2014), (Karmakar and Swarnakar, 2017) and (Gunjan et al., 2012).

1.1 Types of Search Engines

Search engines can be broadly categorized into three types as shown in Figure 1; Crawler-Based Search Engine, Human Powered Search Engine and Hybrid Search Engine. In the crawler-based search engines, listing is created automatically. There is no human intervention involved in the process. Some of the search engines like Google and Yahoo work in this way. They crawl the web going from one web server to the other systematically to create a directory of information in an automated way. Later, when a user submits a query, then that directory

will be consulted to return the results or the matches. If some pages are updated, then the new version of the page will be considered when the crawler visits again.

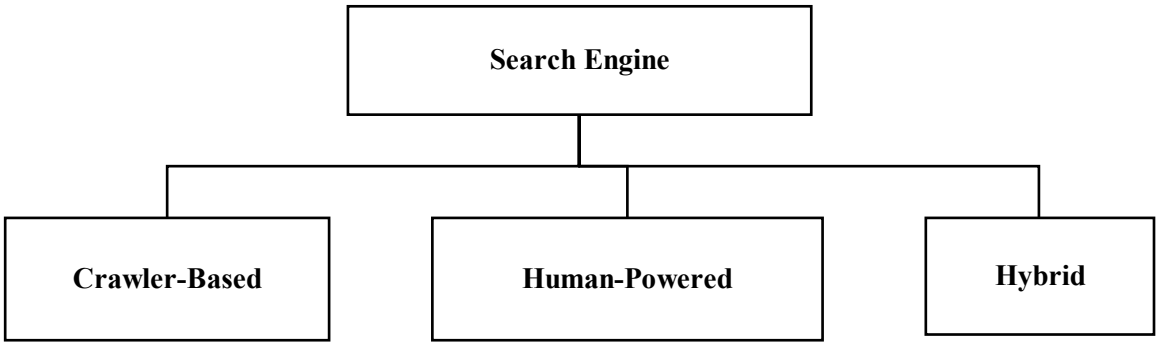

Figure 1: Types of search engine

Human-powered directories rely heavily on humans for its existence. Suppose a user creates a webpage. The user needs to submit a small explanation for the website to be on the list. The short description typically contains keywords and any other information that user want to submit and subsequently when the dictionary is searched in response to some user queries only the descriptions which have been submitted will be searched for. When some changes take place to a page normally, there is no effect on the listing unless the user explicitly informs the directory service that a change has taken place.

Both the approaches have some good features. Crawler-based search engines crawl automatically and in Human-powered search engines creator of the page is submitting the keywords. Best features from both can be combined to get Hybrid search engines. A hybrid search engine can search throughout the directory which is made by web crawlers and also can look into human powered listings. However hybrid search engines must assign some priority.

1.1.1 Components of Crawler-Based Engines

All crawler-based engines have three simple components. These components vary in the way they are implemented and updated. The components of crawler-based search engines are shown in Figure 2.

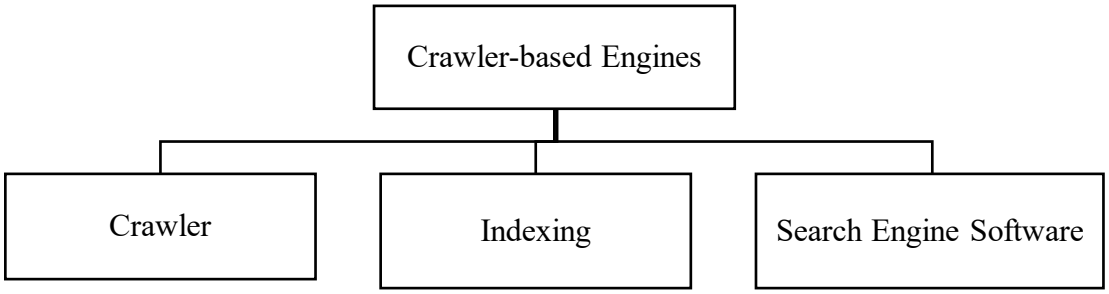

Figure 2: Components of Crawler-Based Search Engines

The first one is called Crawler or Spider. The crawler or spider visits the webpage, recovers the webpage and follow the hyperlinks that lead to other pages within the website. This process is done regularly. Crawling is an important technique as it collects data and keep the data updated. Many researchers like (Singh and Bal, 2014), (Bal and Geetha, 2016), (Amudha, 2017), (Udapure, Kale and Dharmik, 2014) and (Shrivastava, 2018) have analyzed different types of crawling techniques.

The second part is Indexing or Catalog. Indexing means to store and organize the webpages found during the crawling process. The pages which the crawler retrieves are stored in the "indexer" or the "catalogue". It will be updated only when some change to page occurs. Unless a page is not indexed it is not available for search, so indexing is very important.

Third part is software through which the users will submit the queries. This is called "search engine software". It will search through millions of entries in the index and try to find out matches to a search. The search engine software has an additional responsibility to rank the matches using the concept of relevance. Relevance means there should be some mechanism to find out if a specific page is relevant to a given search or not relevant. Based on the relevance, the search results are assigned the ranking. The webpages that are on top are having higher page ranks. This allows the user to get the needed information in the less time. Ranking is an important component of any information retrieval system. Many researchers like (Mohan and Kurmi, 2017), (Tyagi and Sharma, 2012), (Kumar, 2015) and (Kumari, Gupta and Dixit, 2014) have given the comparative study of various algorithms for ranking.

1.1.2 Search Engine Optimization (SEO)

Sorting through several webpages and displaying only the similar webpages is a tough task. Search Engine Optimization (SEO) is a technique that will provide beneficial and quality information out of those several webpages by assigning them ranks. Search engine optimization is an area of research that try to emerge the most relevant result to user. Search engine Optimization techniques can be categorized into two categories- On-page Optimization Techniques and Off-page Optimization Techniques.

On–page Optimization means increasing the rank of a page and improves visitor satisfaction. Off-page Optimization means building back links. Backlinks are the links that point back to the website from other websites. Search engine algorithms give higher ranking to website that has huge number of backlinks. More number of backlinks means increased popularity of website which will lead to increase in the number of users and sessions. Figure 3 listed some of the SEO techniques. On-Page Optimization is done by various techniques given below-

a. Title Optimization: The title of every webpage is optimized so that more and more visitors visit the website. The title of every page should also be assigned uniquely. A search engine uses the title of webpage to determine the relevance of that page.
b. Body text optimization: Using more relevant keywords in the body will lead to higher ranking of webpages and will attract more visitors.
c. Hyperlinks: Efficient hyperlinking between different pages can help a website in securing higher ranking in the search engine results.
d. Meta Tags Optimization: The Meta description tag summarizes a page's content in about 155 characters. This Meta tag description is used by Google to assign the rank to the webpage. So, optimizing the Meta description is crucial for SEO.
e. Sitemaps: A sitemap provides the organization of the site’s content. It enables the search engines to better explore the site, as it will increase the ranking of the webpage.
f. Image optimization: Image optimization is about reducing the file size of the images without losing the quality so that page is loaded in minimum time.

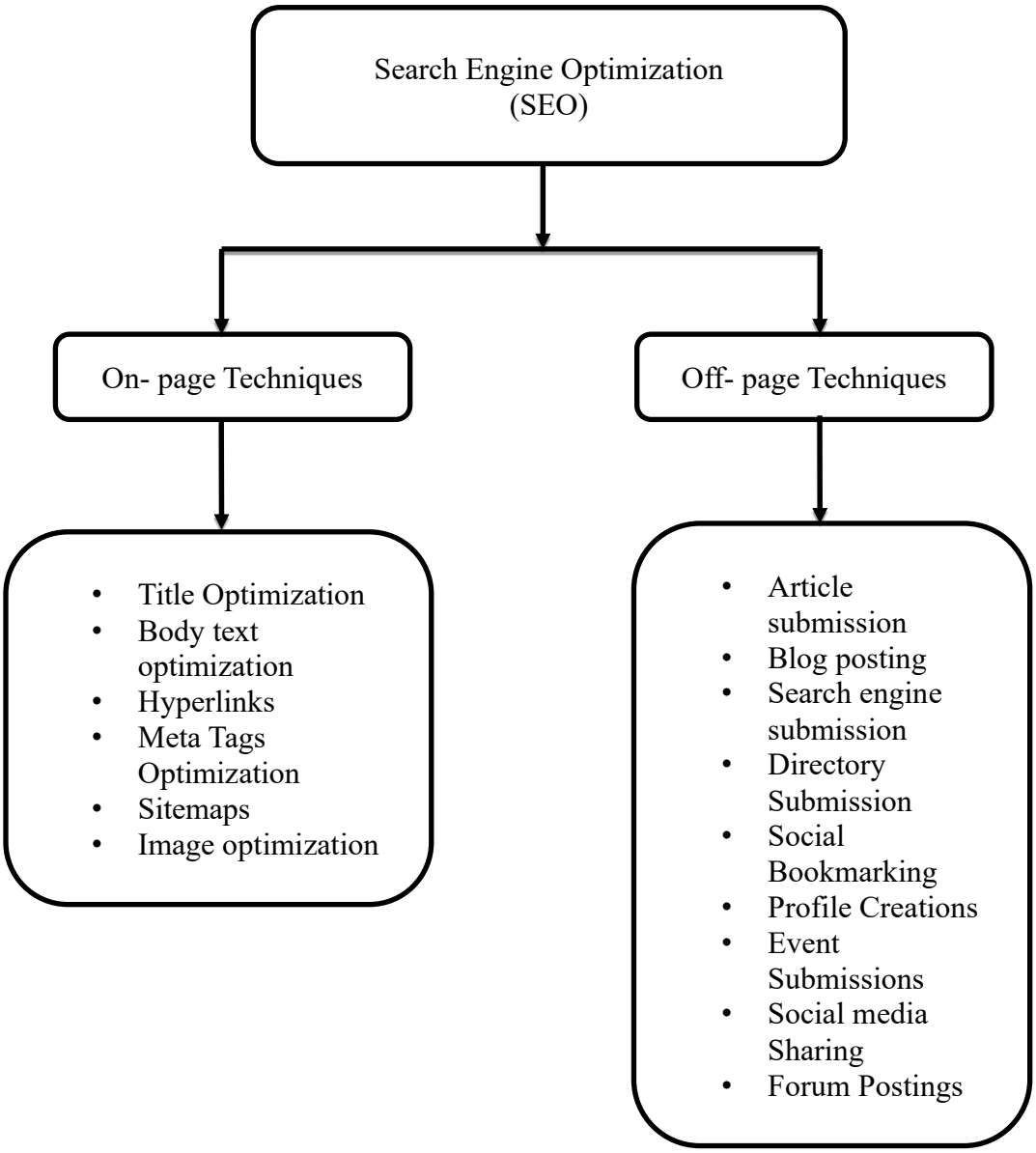

Figure 3: Search Engine Optimization (SEO) Techniques

Off -Page Optimization is done by various techniques explained below:

a. Article submission*:* Article submission means writing articles which are relevant to the website and then submitting them to the popular article submission directories.

b. Blog Posting*:* Blog Posting is also called as the Blog Publishing, it means posting the blog in blogger sites like Blogger, WordPress, Bloglovin, Minds, etc. for getting quality

backlinks. That quality backlinks helps the website to rank higher. Blog posting is done using the selected keywords.

c. Search engine submission: Search engine submission means URL is provided so as to make a search engine aware of a website or page. Adding the URL of the site or page that will be under consideration to a search engine. Website URL are submitted on popular search engine like Google, Yahoo, Altavista, Alltheweb, Lycos, Excite, etc. to get listed for free.
d. Directory Submission*:* Directory Submission means to submit the complete description of the website to popular directory submission sites like Pegasus, Viesearch, etc. The submitted description should contain the link to the website, description of the website, title, etc. The description should be submitted to a category which are related to the keywords.
e. Social Bookmarking*:* Social Bookmarking is a way to store, organize, and manage "bookmarks" of different webpages. It is a technique to promote the website on various social bookmarking sites like Mix, Google Bookmarks, Apsense, Tumblr, Folkd, etc.
f. Profile Creations*:* Profile Creation is used for link building. It is a way to create a visibility and tell others about the website. It is an effective technique to get better back links. It also helps to increase the search engine ranking effectively and receive more users.
g. Event Submissions: Event Submission means to organize an event and post the event details on the event websites. This will create the back link to the website thus generating the users to the website.
h. Social media sharing*:* Social media platforms help to access up-to-date information in every field. It helps to stay in touch with colleagues/friends/relatives and share ideas on various topics. Sharing the information generates the website traffic and increases the visibility of the website.
i. Forum Postings: A forum is a great place to build trust and reputation for the website as a member of that forum. It will ultimately increase the ranking of the website.

Many researchers like (Hussien, 2014), (Kumar and Kumar, 2014), (Omprakash, 2011) and (Singh, 2014) discussed various methods that affect the Search engine optimization and how ranking is also affected by implementing these methods.

1.2 Web Mining

Web Mining is having much importance in Search Engine Optimization (SEO). Using Web mining techniques, the businesses can understand its customers by collecting useful information related to them which will further enhance user experience and customer satisfaction. In Web mining, user needs to pull out data from the content of webpage, hyperlinks of the webpage and server logs maintained at the server. This data is further mined to find relevant information patterns using various data mining techniques. The mined data may have facts that are already in the webpages. It could be the text, structured form of data such as lists and tables, images, video and audio. The goal is to search for patterns in Web-data by assembling and analysing the data so as to realize the trends, the industry and users/visitors. This in turn will help SEO professionals in e-services and e-commerce websites.

1.2.1 Types of Web Mining

Web mining can be categorized into three different types of methods i.e., Web-Content Mining, Web-Structure Mining, and Web-Usage Mining. Figure 4 listed various Web Mining Techniques. The raw data available in Webpages is generally used in Web-Content Mining. The data sources mainly consist of text in webpages like words, tags, etc. Applications of Web-Content Mining are categorization of webpages based on content and ranking of webpages based on content. The structure of different webpages is considered in Web-Structure Mining. The data source mainly used in Web-Structure Mining is the structural information available in Webpage e.g., linking the user to other webpages. Applications of Web-Structure Mining are categorization of Webpages based on linking, Webpages ranking based upon combination of structure and content. Web-Usage Mining is that part of Web Mining which deals with weblog files. The data source mainly consist of the text-logs that are collected when users accesses Web e.g., Common Log Format, Extended Log Format. Applications of Web-Usage Mining are based on modelling techniques, such as user modelling, Web personalization, etc. Web-Usage mining is explained in detail in next section. Many researchers like (Grace, Maheswari and Nagamalai, 2011), (Munk, Kapusta, and Švec, 2010), (Bhavani, Sucharita and

Satyanarana, 2017) and (Das and Turkoglu, 2009) have explained the steps to follow for web mining and how to analyse the log file to get the meaningful information.

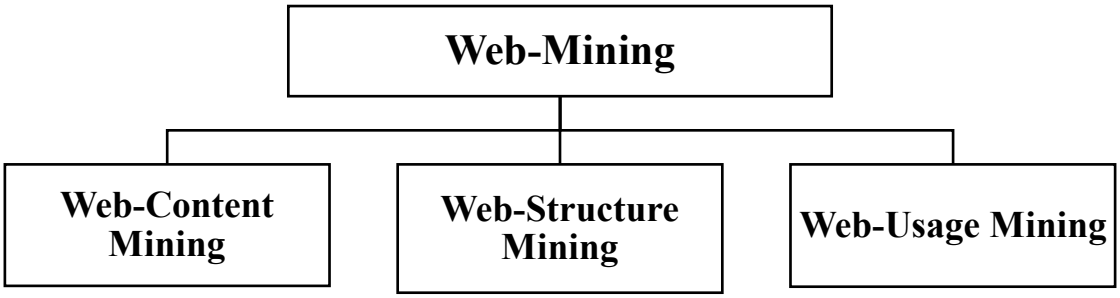

Figure 4: Web Mining Techniques

The power of search engine is enhanced by web mining as it helps in categorizing the webpages and then analysing them. Important advantage of web mining is that searching through the web becomes easy. With the web mining the user behaviour can also be determined. Its highly beneficial for e-commerce websites.

1.2.2 Web-Usage Mining

Web-Usage Mining (WUM) refers to the utilization of Data Mining techniques for getting the meaningful patterns describing the browsing behaviour of users, beginning from access data collected when users are visiting from websites. Web server registers a web log section for every webpage. Web-usage mining is usually used for mining the weblog files and helps to determine the user access patterns of webpages. So, Web-usage mining is also named as log mining. The information can be separated from the weblog file. Only the weblog file is recording the session information about the webpages. Analysis of similarities in weblog records can be beneficial to identify the expected customers for e-commerce companies. Figure 5 gives the overview of steps involved in Web-usage mining.

Following are the steps for web-usage mining –

1. Data collection- Data collection means to collect the log file for knowledge mining and it is the primary task in web mining technique. The data can be gathered from three main sources, which includes (i) web server (ii) proxy server and (iii) web client. The web server information is so voluminous. It contains many detailed information such as date, time in which activities occur, saver's name, IP address, user name, pass-word, required feed, news headlines, and contents. All this information is recorded in the database file.

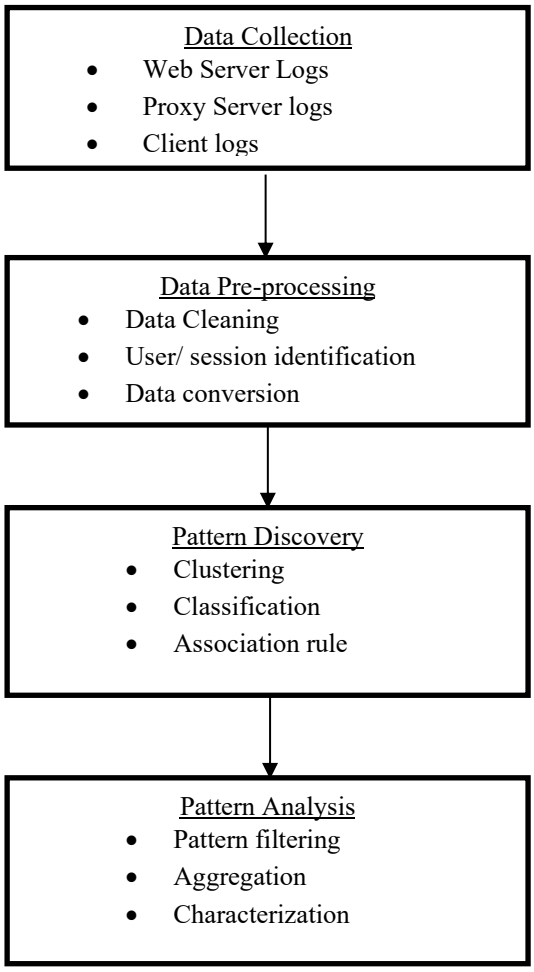

Figure 5: Steps in Web-Usage Mining

2. Data pre-processing- The log file generated by server contains huge volume of data in which most of the instances contains unimportant, incomplete and misleading data. Data pre-processing implies the Data cleaning, session identification and data conversion. Data cleaning is the stage in which insignificant entries are wiped out from the log file. After data cleaning, Session identification will divide the log entries into sessions. Data conversion is to change the format of log file that is required by the knowledge mining algorithms.
3. Pattern discovery- After the data pre-processing phase, the pattern discovery process is performed. In Pattern discovery, groups of users are made. The created groups are having users with same profile and similar search behaviour. The various web-usage mining techniques and algorithms that can be applied for pattern discovery can be path analysis, clustering, Classification and association rule.
4. Pattern analysis-In this last stage of web-usage mining, patterns generated in pattern discovery phase is used for extracting information. The pattern analysis stage transforms the data into useful information.

2. Literature Review

Tyagi and Sharma (2012) proposed a new page ranking mechanism called Weighted PageRank Algorithm based on Visits of Links (VOL), which works on the basis of "Weighted PageRank Algorithm" and takes into account the number of visits of inbound links of webpages.

Singh and Kautish (2015) gave comparison of various page ranking algorithms. They concluded that in the traditional page rank algorithm importance is given to backlinks for calculating the rank of a webpage. In weighted page rank algorithm weight factor is included for both forward and backward links in calculating rank but both the algorithms are calculating ranks at the time of indexing so they are query independent. They concluded that "SimRank Algorithm" ignored the importance of page links and rank is calculated on the basis of content. The results are calculated dynamically at query time. So HITS algorithm computes the rank

considering the forward links, backward links and content but there is a problem of efficiency in this algorithm.

Lu and Cong (2015) proposed a new scheme "Topic_ExpertRank" based on the analysis of links-based and content-based algorithms for improving the accuracy and the efficiency of the search engine.

Hao *et al*. (2015) proposed an improved PageRank Algorithm that judged the relation between different webpages based on their content. The proposed algorithm is showing better precision in searching.

Dutta and Bansal (2017) proposed the improved version for "Weighted PageRank algorithm" based on the efficiency and precision. More relevancies imply more consistency of results. Performance of the proposed algorithm is calculated using Wampserver 2.4.

Yani Achsan and Wibowo (2013) proposed a crawling method that is much faster and cheaper in mining the web database. The proposed crawling method runs on a single system from a single web crawler with hundreds of threads. Later on these threads are distributed among various public servers. The strategy involved in this proposed method leads to improvement in the speed of mining the data and the security of the data is also taken into consideration.

Gupta and Anand (2015) presented various categories of focused web crawler. Web crawler architecture is kept as secret so that search spamming and ranking functions are protected. The proposed architecture of web crawling will expose the secret.

Bal and Geetha (2016) presented client server architecture based "smart distributed crawler" for crawling web. He stated, "In this architecture load between the crawlers is managed by server and each time a crawler is loaded, load is distributed to others by dynamically distributing the URLs". "Focused crawlers" focused on the efficient usage of network bandwidth and storage capacity where as the "Distributed Crawlers" focused on the enhancement of the performance.

Liakos *et al.* (2016) investigated how to build a "Focussed Hidden Web crawler" that can extract topic-specific pages autonomously from the Hidden Web. Conventional web crawlers are not able to discover Hidden web. In order to pull out the terms that capture the significance of the topic they presented an approach that is progressing iteratively and analyses it.

Das and Turkoglu (2009) conducted this study to assist the web designer and web administrator in evaluating the link connections on the website to enhance the impressiveness of a website. Pre-processing of Weblog files is done and the "Path analysis" method is then applied to examine URLs. The methodology proposed is applied to the weblog files of Firat University's web server. The web designer will use the findings and results of this experimental study to redesign the website and make improvement in the website.

Grace, Maheswari and Nagamalai (2011) gave the explanation about the weblog file, its contents, its types and its location, Access Request, IP Address, Bytes Transferred, Time Stamp, etc. They also provided the complete information regarding the preparation of files in web-usage mining process. The various methods that were involved in each step in we-usage mining of the log file is explained in detail with its drawbacks. The concept of extension of logfile is also explained. The entries of extra parameters can be taken into consideration in Log file.

Munk, Kapusta and Švec (2010) described the benefits while analysing the web server log file. In this the visitor does not know that he is the subject for investigation. Disadvantage of this is preparing data from the weblog file, which is the most time-taking task of the analysis of the website. The experiment is performed with the goal to assess the impact of rebuilding the activities of the visitor on the given rules which represent the visitor's behaviour patterns.

Bhavani, Sucharita and Satyanarana (2017) explained that the Web mining can be categorized into three expansive zones of mining i.e., Web-Content Mining, Web-Structure Mining and Web-Usage Mining. While the client is surfing the web, web-usage mining uses the techniques which can analyse the user behaviour. This paper gives the explanation about

the procedure followed in web-usage mining and its applications and devices used as a part of web-usage mining.

3. Description of Broad Area

Search Engines is a software program that helps to find information usually available on the World Wide Web (WWW). Page scanning is performed on the basis of specified keywords and the list of pages is returned in which the corresponding keyword has been located. It is a complicated job to search through multiple webpages and view only the correct webpages. Search Engine Optimization (SEO) is a methodology that can provide valuable and effective information from these webpages, by assigning them ranks and fulfilling the user's needs. Web mining is used to enhance customer satisfaction. The method of web mining is close to the process of data-mining. The distinction typically lies in the collection of data. The gathered data is stored in a data-warehouse during the process of data mining. In case of web mining the data is gathered by crawling various webpages. All of this data is contained in log files. As users search for webpages, a number of log files are generated. Weblogs generated in result of crawling presents many challenges like irrelevant data, usage of complex techniques, incomplete data, etc. There is the need to improve upon these challenges to improve users' experiences.

3.1 Research Gaps

The main challenges in information retrieval using web mining are problem of complexity, problem of time, problem of accuracy and problem of missing entries.

i. Weblogs generated in result of crawling, produces irrelevant data, that needs complex and heavy normalization techniques.
ii. Existing techniques used to normalize the data are time inefficient and intricate.
iii. Weblogs are generally based on user’s last visit to the particular website that cannot be used deterministically to forecast the user interest.
iv. Overloaded server sometimes produces incomplete log or cookies.

3.2 Problem Formulation

An efficient data normalization technique is imperative to remove irrelevant data present in log files. An asymptotically efficient mining model is required to reduce the pre-processing time of weblogs. An efficient model is required to get accurate results to improve user experience. Load balancing at server needs improvement to reduce inconsistency.

4. Objectives of the Study

i. To understand various Search Engine Optimization (SEO) Techniques and their components.
ii. To implement various existing SEO techniques for improvement in ranking of webpages and analyze them based on various parameters.
iii. To develop and design an efficient model for information retrieval in search engine to overcome various challenges.
iv. To simulate the proposed model to establish improvement in user experience.
v. To empirically evaluate the proposed model by comparing with the existing ones.

5. Methodology

In this research work, various on-page and off-page search engine optimization techniques are to be applied to the website and then analysed. The log file would be collected and will be cleaned using various heuristic techniques. After cleaning, machine learning techniques would be used to recognize and analyse the pattern. Based on the pattern analyses, a new model would be proposed to improve the user experience and then the efficiency of the proposed model would be evaluated. Flow for the methodology to be adopted is shown in Figure 6-

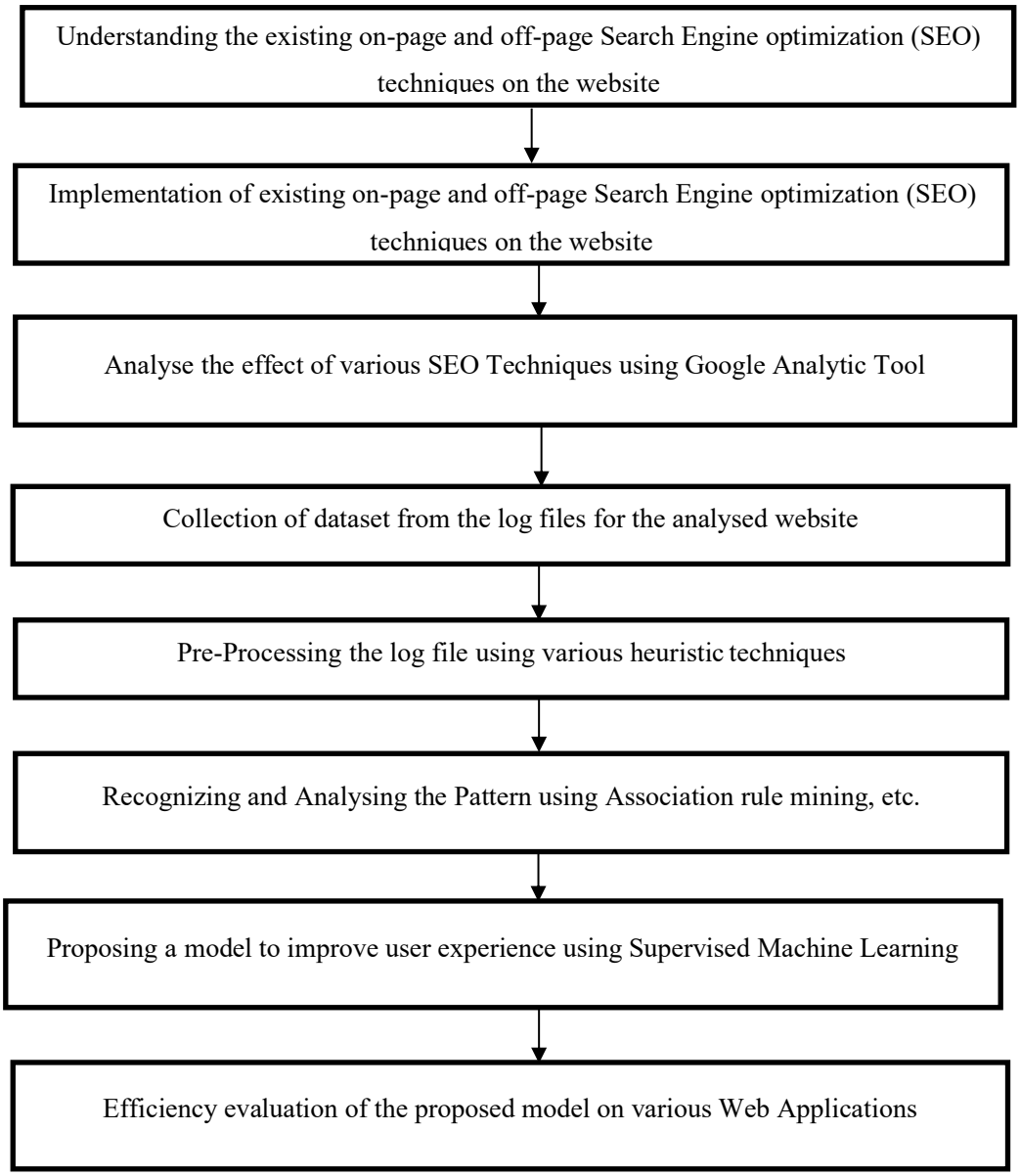
Understanding the existing on-page and off-page Search Engine optimization (SEO) techniques on the website
Implementation of existing on-page and off-page Search Engine optimization (SEO) techniques on the website
Analyse the effect of various SEO Techniques using Google Analytic Tool
Collection of dataset from the log files for the analysed website
Pre-Processing the log file using various heuristic techniques
Recognizing and Analysing the Pattern using Association rule mining, etc.
Proposing a model to improve user experience using Supervised Machine Learning
Efficiency evaluation of the proposed model on various Web Applications

Figure 6: Methodology (Source: Self-generated)

6. Expected Outcome of the Research

Comparison of various Search Engine Optimizations techniques will be done on the website to increase the visibility and ranking of the website. An efficient Information retrieval model will be proposed for predicting the user's behaviour. The analysis of the designed model will be done to verify the improvement during the user's visit to the website.

References

Amudha, S., 2017. Web Crawler for Mining Web Data. *International Research Journal of Engineering and Technology*, *4(2)*, pp.128-136.

Bal, S.K. and Geetha, G., 2016, February. Smart Distributed Web Crawler. In *2016 International Conference on Information Communication and Embedded Systems (ICICES)* (pp. 1-5). IEEE.

Bhavani, B., Sucharita, V. and Satyanarana, K.V.V., 2017. Review on techniques and applications involved in Web Usage Mining. *International Journal of Applied Engineering Research*, *12*(24), pp.15994-15998.

Chen, J.Y., Zheng, H.T., Xiao, X., Sangaiah, A.K., Jiang, Y. and Zhao, C.Z., 2017. Tianji: Implementation of an Efficient tracking engine in the Mobile Internet era. *IEEE Access*, *5*, pp.16592-16600.

Das, R. and Turkoglu, I., 2009. Creating meaningful data from Web Logs for Improving the Impressiveness of a website by using Path Analysis Method. *Expert Systems with Applications*, *36*(3), pp.6635-6644.

Dubey, H. and Roy, B.N., 2011. An Improved Page Rank Algorithm based on Optimized Normalization Technique, *International Journal of Computer Science and Information Technologies*, 2(5), pp.2183-2188.

Dutta, M. and Bansal, K.L., 2017. Improvement in Weighted Page Rank Algorithm using Efficiency and Precision, *International Journal of Computer Science & Engineering Technology (IJCSET)*, 8(1), pp.11-24.

Grace, L.K., Maheswari, V. and Nagamalai, D., 2011. Analysis of Web Logs and Web User in Web Mining. *International Journal of Network Security & Its Applications (IJNSA),* 3(1), pp.99-110.

Gupta, A. and Anand, P., 2015, February. Focused Web Crawlers and its approaches. In *2015 International Conference on Futuristic Trends on Computational Analysis and Knowledge Management (ABLAZE)* (pp. 619-622). IEEE.

Gunjan, V.K., Kumari, M., Kumar, A. and Rao, A.A., 2012. Search Engine Optimization with Google. *International Journal of Computer Science Issues (IJCSI)*, *9*(1), pp.206-214.

Hao, Z., Qiumei, P., Hong, Z. and Zhihao, S., 2015, August. An improved PageRank algorithm based on Web Content. In *2015 14th International Symposium on Distributed Computing and Applications for Business Engineering and Science (DCABES)* (pp. 284-287). IEEE.

Hussien, A.S., 2014. Factors affect Search Engine Optimization. *International Journal of Computer Science and Network Security*, *14*(9), pp.28-33.

Karmakar, S. and Swarnakar, S., 2017. New Concept based Indexing Technique for Search Engine. *Indian Journal of Science and Technology*, *10*(18), pp.1-10.

Kumar, D., 2015. A Review Paper on Page Ranking Algorithms. *International Journal of Advanced Research in Computer Engineering & Technology (IJARCET)*, *4*(6), pp.2806-2811.

Kumar, K. and Mukoko, F.D., 2013. PageRank Algorithm and its variations: A Survey report. *IOSR Journal of Computer Engineering (IOSR-JCE)*, *14*(1), pp.38-45.

Kumar, L. and Kumar, N., 2014. SEO Techniques for a Website and its Effectiveness in Context of Google Search Engine. *International Journal of Computer Sciences and Engineering (IJCSE)*, *2(4)*, pp.113-118.

Kumari, T., Gupta, A. and Dixit, A., 2014. Comparative study of Page Rank and Weighted Page Rank Algorithm. *International Journal of Innovative Research in Computer and Communication Engineering*, *2*(2), pp.2929-2937.

Liakos, P., Ntoulas, A., Labrinidis, A. and Delis, A., 2016. Focused Crawling for the hidden web. *World Wide Web*, *19*(4), pp.605-631.

Liu, D., Xu, X. and Long, Y., 2017, May. On member Search Engine selection using Artificial Neural Network in meta Search Engine. In *2017 IEEE/ACIS 16th International Conference on Computer and Information Science (ICIS)* (pp. 865-868). IEEE.

Lu, H., Zhan, D., Zhou, L. and He, D., 2016. An improved Focused Crawler: using Web Page Classification and Link Priority Evaluation. *Mathematical Problems in Engineering*, *2016*, pp.1-10.

Lu, P. and Cong, X., 2015. The Research on Webpage Ranking Algorithm based on Topic-Expert Documents. In *Recent Advances in Information and Communication Technology 2015* (pp. 195-204). Springer, Cham.

Mohan, K. and Kurmi, J., 2017. A Technique to Improved Page Rank Algorithm in perspective to Optimized Normalization Technique. *International Journal of Advanced Research in Computer Science*, *8*(3) , pp.702-706.

Munk, M., Kapusta, J. and Švec, P., 2010. Data Preprocessing Evaluation for Web Log Mining: reconstruction of activities of a web visitor. *Procedia Computer Science*, *1*(1), pp.2273-2280.

Omprakash, K.S.V., 2011. Concept of Search Engine Optimization in Web Search Engine. *International Journal of Advanced Engineering Research and Studies*, *1*(1), pp.235-237.

Pranav, A. and Chauhan, S., 2015. Efficient focused Web Crawling approach for Search Engine. *International Journal of Computer Science and Mobile Computing*, *4*(5), pp.545-551.

Rani, R. and Jain, V., 2013. Weighted PageRank using the Rank Improvement. *International Journal of Scientific and Research Publications*, *3*(7), pp.1-5.

Rohatgi, S. and Kundu, S., 2017. Focused Crawling Using Content Classification and Link Priority Estimation. *International Journal of Scientific & Engineering Research*, 8(10), pp.1223-1228.

Sahu, M.B. and Bharne, S., 2016. A Survey on Various Kinds of Web Crawlers and Intelligent Crawler. *International Journal of Scientific Engineering and Applied Science (IJSEAS)*, *2(3),* pp.468-472.

Shrivastava, V., 2018. A methodical study of Web Crawler. *International Journal of Engineering Research and Application*, *8*(11), pp.01-08.

Singh, C. and Kautish, S.K., 2015. Page Ranking Algorithms for Web Mining: A Review. *International Journal of Computer Applications*, pp.17-21.

Singh, M.S. and Bal, J.S., 2014. Web Crawler: Extracting the Web Data. *International Journal of Computer Trends and Technology*, *13*(3), pp.132-137.

Singh, P., 2014. A Review Paper on SEO based Ranking of Web Documents. *International Journal of Advanced Research in Computer Science and Software Engineering*, *4*(7) , pp.1136-1140.

Tyagi, N. and Sharma, S., 2012. Comparative study of various Page Ranking Algorithms in Web Structure Mining (wsm). *International Journal of Innovative Technology and Exploring Engineering (IJITEE)*, *1*(1), pp.14-19.

Tyagi, N. and Sharma, S., 2012. Weighted Page Rank Algorithm based on number of visits of links of Web page. *International Journal of Soft Computing and Engineering (IJSCE) ISSN*, 2(3), pp. 441-446.

Udapure, T.V., Kale, R.D. and Dharmik, R.C., 2014. Study of Web Crawler and its different types. *IOSR Journal of Computer Engineering*, *16*(1), pp.01-05.

Wen, B., 2018. Application of Distributed Web Crawlers in Information Management System. *Informatica*, *42*(1), pp.137–143.

Yani Achsan, H.T. and Wibowo, W.C., 2013. A Fast Distributed Focused-Web Crawling. *Annals of DAAAM & Proceedings*, *24*(1), pp.492–499.

MANAV RACHNA
vidyanatariksha

Dr. O.P. Bhalla Central Library
Manav Rachna International Institute of Research and Studies
(A deemed-to-be-university under section 3 of UGC Act 1956)

SIMILARITY VERIFICATION CERTIFICATE

Title: An EFFICIENT WEB INFORMATION RETRIEVAL MODEL USING MACHINE LEARNING

Researcher: SIMPLE SHARMA

Registration Number: 18/Ph.D./106

Supervisor: DR. SEEMA VERMA

Department: C.S.E

This is to certify that the above mentioned Short ~~Synopsis~~/Thesis was scanned for similarity index through Turnitin Software on 23/3/2021 with submission ID 1542776902 and the similarity index was found to be 10 %. The complete report running into 23 Pages, 5189 Word count with File Size 288 KB/~~MB~~ is being forwarded to the Research Supervisor/DRC/DRB for review.

Date: 23/3/2021

Dr. Rajesh Kumar
University Librarian
Dr. O.P. Bhalla Central Library
Manav Rachna International
Institute of Research & Studies
Faridabad

The complete report of the above Short Synopsis/Thesis has been reviewed and the similarity index is below the accepted norms as per the University Grants Commission (Promotion of Academic Integrity and Prevention of Plagiarism in Higher Educational Institutions) Regulations 2018. The Short Synopsis/Thesis may be considered for further processing

Research Scholar

Research Supervisor

Short Synopsis Simple Sharma 26.03.2021

by Simple 4 Sharma

Submission date: 26-Mar-2021 11:25AM (UTC+0530)
Submission ID: 1542776902
File name: Simple_Sharma.doc (288K)
Word count: 5189
Character count: 28530

An Efficient Web Information Retrieval Model using Machine Learning

SHORT SYNOPSIS

By

Ms. Simple Sharma
Registration number: 18/ Ph.D./106

Under the supervision of
Dr. Seema Verma

ABSTRACT

Search Engines is a software program that looks for items and recognizes those items in a database that makes a perfect match to keywords or characters identified by the users, used for finding a particular website on the World Wide Web uniquely. Searching is done with given keywords and a list of webpage documents are reflected based on the matched keywords. Sorting and selection through various webpages is a tedious task in search engine. Search Engine Optimization (SEO) is a method that provides useful information out of billions of webpages with the use of various techniques. Search engines are broadly divided into three categories namely; Crawler-Based Search Engines, Human-Powered Directories and Hybrid Search Engine. Information retrieval in Search Engine is a big challenge. Web mining helps to boost the power of search engine by searching and analysing patterns in web data. Web mining helps SEO professional to analyse information and trends. The mined data can be the text, structured form of data such as lists and tables, images, video and audio. Web mining results in various complicacies while analysing the information pattern. The foremost challenge is the generation of irrelevant data in weblogs due to which complex pre-processing techniques are in use. The complex techniques used in pre-processing consume more time to get the formatted output; hence time delay is the other challenge. Another big challenge is inaccuracy due to previous information stored in the weblogs. The weblogs contain old information based on the user's previous visit to the site, which affects the

current requirements. Hence, there is the need to analyse the weblogs for various challenges to optimize the information retrieval from search engines. This research aims to suggest an efficient information retrieval model to improve users' experience after analysing various complexities.

Key words: Search Engine, Crawler, Page-rank, Search Engine Optimization, Relevancy, Web Mining, Web-Usage Mining, Ranking.

1. Introduction

Search engine is a software program that helps in searching information on the World Wide Web. Searching through billions of webpages to get the required page is a tedious task. Some of the popular search engines are Google, Yahoo, Bing, Ask, etc. Search engines have to call through all the websites to carry out indexing. They may be slow to index dynamically because the contents are not static. The speed at which the search engine is indexing the pages is increasing day by day but the documents over the Web are increasing at a much faster rate. There are billions of pages that exist in the World Wide Web (WWW). Search engines create a dictionary where size of the dictionary needs to be very large. Today search engines are becoming more efficient and powerful by using various algorithms and techniques to provide the best results demanded by user. As an example, Google has introduced some simple concepts which allows a user to get more relevant search results as compared to other competing search engines like, link popularity, the number of pages on a website, recording pages that point to another webpage, etc. All these concepts increased the visibility of the page and the website effectively communicates with major search engines. Optimizing a page involves editing the contents of the webpage so as to extend its relevancy to identified words. Many researchers have developed numerous search engine optimization strategies, including (Gupta and Anand, 2015), (Liu, Xu and Long, 2017), (Omprakash, 2011), (Chen *et al.,* 2017), (Kumar and Kumar, 2014), (Singh, 2014), (Karmakar and Swarnakar, 2017) and (Gunjan et al., 2012).

1.1 Types of Search Engines

Search engines can be broadly categorized into three types as shown in Figure 1; Crawler-Based Search Engine, Human Powered Search Engine and Hybrid Search Engine. In the crawler-based search engines, listing is created automatically. There is no human intervention involved in the process. Some of the search engines like Google and Yahoo work in this way. They crawl the web going from one web server to the other systematically to create a directory of information in an automated way. Later, when a user submits a query, then that directory will be consulted to return the results or the matches. If some pages are updated, then the new version of the page will be considered when the crawler visits again.

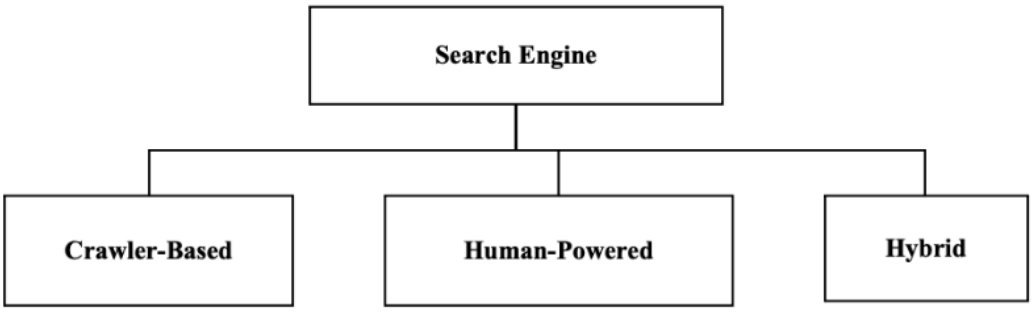

Figure 1: Types of search engine

Human-powered directories rely heavily on humans for its existence. Suppose a user creates a webpage. The user needs to submit a small explanation for the website to be on the list. The short description typically contains keywords and any other information that user want to submit and subsequently when the dictionary is searched in response to some user queries only the descriptions which have been submitted will be searched for. When some changes take place to a page normally, there is no effect on the listing unless the user explicitly informs the directory service that a change has taken place.

Both the approaches have some good features. Crawler-based search engines crawl automatically and in Human-powered search engines creator of the page is submitting the keywords. Best features from both can be combined to get Hybrid search engines. A hybrid search engine can search throughout the directory which is made by web crawlers and also can look into human powered listings. However hybrid search engines must assign some priority.

1.1.1 Components of Crawler-Based Engines

All crawler-based engines have three simple components. These components vary in the way they are implemented and updated. The components of crawler-based search engines are shown in Figure 2.

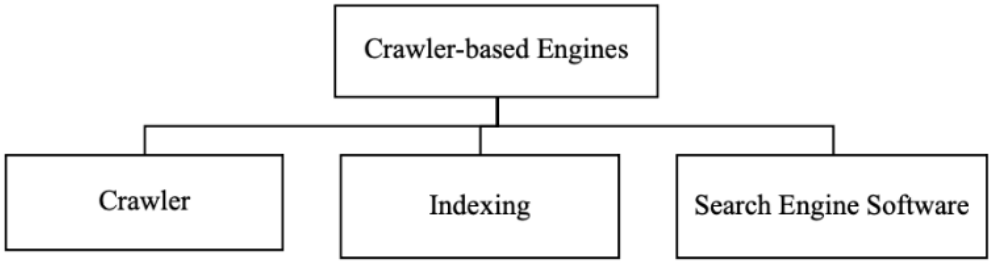

Figure 2: Components of Crawler-Based Search Engines

The first one is called Crawler or Spider. The crawler or spider visits the webpage, recovers the webpage and follow the hyperlinks that lead to other pages within the website. This process is done regularly. Crawling is an important technique as it collects data and keep the data updated. Many researchers like (Singh and Bal, 2014), (Bal and Geetha, 2016), (Amudha, 2017), (Udapure, Kale and Dharmik, 2014) and (Shrivastava, 2018) have analyzed different types of crawling techniques.

The second part is Indexing or Catalog. Indexing means to store and organize the webpages found during the crawling process. The pages which the crawler retrieves are stored in the "indexer" or the "catalogue". It will be updated only when some change to page occurs. Unless a page is not indexed it is not available for search, so indexing is very important.

Third part is software through which the users will submit the queries. This is called "search engine software". It will search through millions of entries in the index and try to find out matches to a search. The search engine software has an additional responsibility to rank the matches using the concept of relevance. Relevance means there should be some mechanism to find out if a specific page is relevant to a given search or not relevant. Based on the relevance, the search results are assigned the ranking. The webpages that are on top are

Short Synopsis Simple Sharma 26.03.2021

Dr. Rajesh Kumar
University Librarian
Dr. O.P. Bhalla Central Library
Manav Rachna International
Institute of Research & Studies
Faridabad

Excluding matches
less than 14 words

ORIGINALITY REPORT

10%	7%	6%	5%
SIMILARITY INDEX	INTERNET SOURCES	PUBLICATIONS	STUDENT PAPERS

PRIMARY SOURCES

1	www.ijritcc.org Internet Source	1%
2	Facca, F.M.. "Mining interesting knowledge from weblogs: a survey", Data & Knowledge Engineering, 200506 Publication	1%
3	Submitted to Middlesex University Student Paper	1%
4	docplayer.net Internet Source	1%
5	keddiyan.com Internet Source	1%
6	krishikosh.egranth.ac.in Internet Source	1%
7	"Proceedings of International Conference on Computational Intelligence and Data Engineering", Springer Science and Business Media LLC, 2018 Publication	1%

8	Submitted to Pacific University Student Paper	1%
9	Submitted to Vels University Student Paper	1%
10	Submitted to South Dakota Board of Regents Student Paper	<1%
11	link.springer.com Internet Source	<1%
12	www.ijert.org Internet Source	<1%
13	Das, R.. "Creating meaningful data from web logs for improving the impressiveness of a website by using path analysis method", Expert Systems With Applications, 200904 Publication	<1%
14	Submitted to Higher Education Commission Pakistan Student Paper	<1%
15	ieeexplore.ieee.org Internet Source	<1%
16	Submitted to University of Portsmouth Student Paper	<1%
17	Submitted to University of Pune Student Paper	<1%

18	Shahnaz Parvin Nina. "Pattern Discovery of Web Usage Mining", 2009 International Conference on Computer Technology and Development, 11/2009 Publication	<1%
19	andrews.gknu.com Internet Source	<1%

Exclude quotes	On	Exclude matches	< 14 words
Exclude bibliography	On		